ED 419 817 TM 028 392

AUTHOR Mundfrom, Daniel J.; Whitcomb, Alan

TITLE Imputing Missing Values: The Effect on the Accuracy of Classification.

PUB DATE 1998-04-15

NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Classification; Heart Disorders; *Patients; Predictor Variables; Regression (Statistics)

IDENTIFIERS *Accuracy; Hot Deck Procedures; Imputation; *Missing Data

ABSTRACT

Data from records of 99 patients were used to classify cardiac patients as to whether they were likely or unlikely to experience a subsequent morbid event after admission to a hospital. Both a linear discriminant function and a logistic regression equation were developed using a set of nine predictor variables that were chosen on the basis of their correlations with the likelihood of a subsequent morbid event. Once the models were obtained, artificially-generated missing values were replaced with imputed values using mean substitution, regression imputation, and hot-deck imputation techniques. The effect on the accuracy of the predictions using models with imputed values was determined by comparing the reclassifications using imputed data with the actual occurrence or nonoccurrence of a subsequent morbid event. Mean substitution and hot-deck imputation performed slightly better than regression imputation in this application regardless of whether or not the predictor variable whose values were being imputed was categorical or numerical. (Contains 2 tables and 18 references.) (Author/SLD)

Imputing Missing Values:
The Effect on the Accuracy of Classification

Daniel J. Mundfrom and Alan Whitcomb
University of Northern Colorado

A paper presented at the American Educational Research Association Annual Meeting
San Diego, CA, April 15, 1998

Imputing Missing Values:
The Effect on the Accuracy of Classification

Abstract

Data from patient records were used to classify cardiac patients as to whether they are likely or unlikely to experience a subsequent morbid event after admission to a hospital. Both a linear discriminant function and a logistic regression equation were developed using a set of nine predictor variables which were chosen on the basis of their correlations with the likelihood of a subsequent morbid event. Once the models were obtained, artificially-generated missing values were replaced with imputed values using mean substitution, regression imputation and hot-deck imputation techniques. The effect on the accuracy of the predictions using models with imputed values was determined by comparing the re-classifications using imputed data with the actual occurrence or non-occurrence of a subsequent morbid event. Mean substitution and hot-deck imputation performed slightly better than regression imputation in this application regardless of whether or not the predictor variable whose values were being imputed was categorical or numerical.

Imputing Missing Values: The Effect on the Accuracy of Classification

Introduction

Statistical modeling techniques have been widely used for many years to predict a particular outcome using information from a group of variables which are related to the outcome of interest. That outcome could be a continuous variable such as an achievement test score or a categorical variable such as whether or not an individual graduated from a particular graduate program. When the outcome of interest is continuous, the appropriate statistical procedures would generally be a multiple regression analysis or an analysis of variance. When the outcome of interest is dichotomous, the analysis reduces to classifying an individual into one of two or more groups depending on the observed values of a set of predictor variables, and the appropriate procedure to use is either a discriminant analysis or logistic regression.

One situation in which statistical modeling, with a dichotomous outcome variable, could be used for classification involves the decision that rural hospitals must make when a cardiac patient presents at the hospital. Rural hospitals frequently cannot afford all the latest technological equipment that their larger urban counterparts have available. One possible decision would be to automatically send all cardiac patients on to the urban hospital. This decision has obvious benefits, but also has at least two drawbacks. One drawback is that some patients will be sent who could have been cared for sufficiently in the local hospital. This decision requires needless expense for the patient that could have been avoided. A second drawback is that every patient transported away from the local hospital takes with him/her revenue that could have been spent locally that would help the local hospital maintain economic viability.

Another decision that could be made is to keep all patients and care for them locally. While this decision keeps the revenue "at home," it may not be in the best interest of every patient in terms of providing them with the necessary care. The desire to balance the patients' needs for having the best possible care and the hospitals' needs to maintain their economic vitality forms the framework for this research.

One way to try and balance these needs is to reduce the number of unnecessary patient transportations from rural hospitals to tertiary care facilities. Technology to assist the rural physician in more accurately predicting which cardiac patients are likely to experience a morbid event and which are not can reduce the number of patient transportations. Since cardiovascular disease is the leading cause of death in the United States, and its prevalence is highest in rural areas where the latest advancements in providing necessary care may not be available, predicting likely candidates for a subsequent morbid event would be a valuable asset for the rural physician. Coronary Care Units (CCU's) have proven to be extremely effective in preventing death from certain cardiac events, but the cost of these units normally limits their presence to tertiary care facilities. Moreover, predicting which cardiac patients are likely to experience a serious morbid event has proven difficult, with only about 25% of patients in a CCU suffering life-threatening events during their stay. The ability to make an accurate prediction would increase the economic viability of the rural hospital and also reduce the financial burden for the patient, without having a negative impact on the adequacy of the care the patient received.

Statistical models, based on patient data collected at the time of the initial hospital visit, could be useful for making cardiac morbidity predictions. However, it is not always possible to obtain a measurement on every variable of interest in real data situations. Missing data values have plagued statisticians for years in their attempts to obtain useful, accurate summaries and predictions. Missing data is an even greater concern when decisions must be made about the appropriateness of the care a patient should receive. From a methodological perspective, missing values either reduce the number of available cases for analysis or introduce bias into the estimation and/or prediction process. Neither scenario is desirable.

The purpose of this research was twofold. The first objective was to develop a statistical model that could be used to predict which cardiac patients are likely to experience a subsequent The model that was developed was based upon the complete-case data of actual cardiac patients.

The second objective was to examine the effects on the accuracy of the model's predictions when imputed values from three different imputation techniques were substituted for artificially-generated missing data. Of particular interest was determining which of the techniques would have the smallest detrimental effect on the accuracy of predictions when using imputed values.

Background

The initial phase of this research was to select a suitable model for predicting a morbid event in cardiac patients. Five morbid events were identified and defined as follows: development of sustained ventricular tachycardia, ventricular fibrillation, cardiogenic shock, development of myocardial infarction or extension of infarction, and bradycardia of less than 45 beats per minute. Identifying potential predictor variables that may be indicators of one or more of these events was the next step.

Although prevention of fibrillation has long been recognized as desirable (Lown, Fakhro, Hood & Thorn, 1967), defining specific electrical parameters heralding fibrillation has not been easy (Campbell, Murray & Julian, 1981). Like ventricular fibrillation, predicting the development or worsening of pump failure has also been difficult. Nonetheless, numerous studies now exist that have attempted to accurately define the clinical predictors associated with a poor prognosis in CCU patients. Parameters as varied as age, hypertension, diabetes, length of stay in the CCU (Gheorghiade, et. al., 1987), ST and T wave changes (Severi, et. al., 1988; Bell, Montarello & Steele, 1990), sex, anterior infarction, hypotension at admission, ventricular tachyarrhythmias, diabetes, Killip class III and IV (De Martini, et. al., 1990), previous myocardial infarction (Nishi, et. al., 1992), and serum urea (Marik, Lipman, Eidelman & Erskine, 1990) have all been shown to have short-term prognostic significance.

Assuming that a set of suitable variables for predicting a morbid event can be identified, the problem of missing data must still be addressed. In many real-life situations, one or more of the individual cases will have incomplete data. In this application, one or more of the signs necessary for optimally predicting a morbid event may be unavailable. Perhaps a measurement goes unrecorded, a test is not available to be run, or the results of a test are inadvertently lost. Most standard statistical techniques build their models using only those cases which have a complete set of data values. If the value for even one variable is missing, the entire set of measurements for that individual is excluded from the model-building process. Complete-case analyses are often used because of simplicity of analysis and for comparability resulting from using a common sample for all calculations (Little & Rubin, 1987). However, the loss of potentially useful information in the data which is discarded is undesirable.

Another problem occurs if, after the model has been obtained, one or more of the values required to use the model are unavailable. The optimal model is constructed based on the assumption that data will be available for each variable included in the model. A regression coefficient is calculated for each variable, so its contribution to the prediction of the outcome variable is appropriately weighted. If even one value is missing for an individual, the optimal model cannot be used appropriately and if it is used anyway, the resulting prediction may be suspect. The problem of missing data can be overcome by deleting cases with missing values or by replacing missing values with an imputed value. Imputed values are generally obtained from the existing data and there are a variety of techniques available for imputation, each having different properties that make them more or less useful in any particular situation (see Buck, 1960; Affifi & Elasahoff, 1966, Haitovsky, 1968; Hartley & Hocking, 1971; Chan & Dunn, 1972; Rubin, 1976; Little & Rubin, 1987; Rubin, 1991; Rindskopf, 1992; van Buuren & van Ruckevorsel, 1992; Kromrey & Hines, 1994; Roth & Switzer, 1995).

Among the most commonly recommended missing data treatments are listwise and pairwise deletion, mean substitution, regression imputation, hot-deck imputation, and the EM algorithm (Little & Rubin, 1987). The selection of which of these procedures to examine in this research involves several considerations. First of all, the deletion techniques were deemed inappropriate, since in this phase of this research, a model is not being constructed, but rather a previously built model will be used to classify an individual. Deleting the case would result in the lack of such classification for that individual, an outcome that is unacceptable in this situation. Hence, only the imputation techniques received further consideration. Accuracy of classification is the primary issue, but the principal purpose of this research is to compare accuracy rates, so this characteristic was not used to select techniques for consideration. Ease of use is also a primary factor. In the context of predicting a morbid event, the desire was to use an imputation technique that would not require the physician to perform a complicated or time-consuming task in order to make a decision regarding a particular patient. The need to keep this part of the process simple therefore became the primary criteria for selecting an imputation technique.

Another consideration was the task at hand. The value being predicted was the likelihood of a morbid event. Because this is basically a classification problem (classifying an individual into one of two groups; likely to have a subsequent morbid event or unlikely to have one), an imputation technique with optimal properties in discrimination was desired. Chan & Dunn (1972) reported that mean substitution and the principal components method outperformed other techniques for classification. Kim & Curry (1977) and Raymond & Roberts (1987) report that regression imputation has the desirable property of minimizing the variability in the imputed values. Hot-deck imputation is frequently used in practice because of its intuitive appeal (Roth & Switzer, 1995), but little research regarding its accuracy has been done. Rubin (1991) lists several desirable properties of the EM algorithm that seem to indicate it as the procedure of choice in many situations, especially with large samples. All four of these imputation techniques have desirable characteristics.

From this list of four techniques, the EM algorithm, although highly regarded for many reasons, was deemed to be too complex to have a reasonable expectation of use by a physician in practice. Consequently, the techniques chosen to be examined in this research were mean substitution, regression imputation, and hot-deck imputation. Recognizing that this decision is subjective and may not necessarily be optimal, it still seemed reasonable that due to the relative simplicity of using these procedures, that if any were found to be sufficiently accurate, it would have a high expectation for use in practice.

Method

The archival data used in this research were obtained from patient records for a sample of 99 cardiac patients who had been admitted over a three-year period to a Cardiac Care Unit or a Cardiac Monitored Care Unit (MCU) in an urban University-affiliated hospital after suffering a morbid event for which data existed on a list of 29 variables which had been identified as potential predictors of a subsequent morbid event after suffering an initial such event. Patients who had undergone surgery in the six month period prior to admittance to the CCU/MCU or who were on mechanical breathing support were excluded from the sample. In this sample, 38 individuals experienced at least one subsequent morbid event in the hospital after being admitted.

This list of variables included the continuous variables: height, weight, age, systolic blood pressure, diastolic blood pressure, hematocrit, serum potassium level, serum creatine level, white blood cell count, respiration rate, and heart rate, and the categorical variables: sex, current myocardial infarction, evidence of anterior infarction, atrial arrhythmia, ventricular arrhythmia, S-T depression, diabetes, previous infarction, smoking, rales greater than 1/3 up, presence of heart sound S3, syncope, ventricular ectopics, use of aspirin in treatment, and use of Class I, II, III, or IV drugs. This initial list of potential predictors was reduced from 29 to 9 based upon their correlations with the occurrence or non-occurrence of one or more of the five morbid events ($|r| > .1643$). The final group of nine predictors included sex, age, weight, systolic blood pressure, white blood cell count, ventricular arrhythmia (an indicator of abnormal heart rhythm; measured as present or absent), syncope (an indicator of poorly oxygenated blood; measured as poor or not poor), heart sound S3 (an indicator of heart valve insufficiency; measured as sufficient or not), and use of aspirin (measured as used in treatment or not).

Once these predictors were identified, a linear discriminant function, based on those nine predictors, was created for classifying patients as likely to experience a subsequent morbid event or not likely to experience such an event. Similarly, a logistic regression equation using the same set of nine predictors was generated for the same classification purpose. The number of correct classifications for each model was determined by comparing classifications resulting from use of the statistical model with the actual occurrence or non-occurrence of subsequent morbid events.

To investigate the effect of different techniques for imputing missing values, values for one predictor at a time were deleted for each of the 99 patients and replaced with an imputed value. After replacing the original value with an imputed value, the number of correct re-classifications using the original discriminant function and the logistic regression model were calculated. In turn, this process was repeated for each of the three imputation techniques and for eight of the nine predictor variables. (It was decided that the variable sex is unlikely to ever be unknown in this context, so replacing the actual value of the sex variable with an imputed value seemed unnecessary.) The number of correct re-classifications, using imputed values in both the discriminant analysis and the logistic regression analysis, were then compared to the number of correct classifications using the original data.

For the mean imputation technique, imputed values for a particular variable were obtained by calculating the mean value for that predictor using all 99 patients' records. Using a single variable at a time for imputation, the original values of that variable were replaced with the mean value of that variable in each of the individuals' records. The other eight predictors were left unchanged and the individual was re-classified into one of the two groups. The value for each of the other predictors, excluding sex, was replaced with its mean value in the same way, each time using the original data values for the other predictors, and each individual was re-classified.

Using the regression imputation technique, imputed values for each predictor were calculated for the patients by building a regression equation involving the other eight predictors. Imputed values for the variables which were measured on a continuous scale (e.g., age) were determined using multiple linear regression analysis. For the dichotomous predictors (e.g., ventricular arrhythmia), a logistic regression analysis was used to build the model for prediction. The coefficients used to generate the imputed values for each of the eight predictors (again, excluding sex) are presented in Table 1.

Table 1. Coefficients of Predictor Variables Used in Regression Imputation

					Response Variable			
	Age	Weight	SBP	WBCC	VA	Syncope	S3	Aspirin
Coefficients	(MLR)	(MLR)	(MLR)	(MLR)	(LR)	(LR)	(LR)	(LR)
Constant	56.005	82.124	105.324	8.302	0.334	2.463	-1.523	0.062
Gender	-0.481	-7.363	9.060	0.975	-0.247	0.820	1.330	0.407
Age		-0.404	0.417	0.030	-0.016	-0.037	-0.008	-0.020
Weight	-0.209		0.262	0.038	-0.001	-0.0001	0.019	-0.011
SBP	0.106	0.129		-0.029	0.013	0.008	0.014	0.012
WBCC	0.311	0.759	-1.165		0.018	-0.004	-0.074	0.127
VA	2.853	0.164	-8.705	-0.289		-0.473	-1.291	-0.454
Syncope	6.973	-0.100	-7.218	-0.061	-0.445		2.298	0.585
S3	1.057	-5.948	-9.849	1.732	-1.036	2.151		1.869
Aspirin	3.589	3.131	-9.036	-1.181	-0.322	0.528	2.260	

Note: SBP = systolic blood pressure; WBCC = white blood cell count; VA = ventricular arrhythmia;; MLR = multiple linear regression; LR = logistic regression

Using the hot-deck imputation technique, an imputed value for one predictor was obtained for each patient by randomly selecting (with replacement) a value from that variable's original set of 99 values. However, since the randomly selected values would vary from one selection to another, so would the number of correct re-classifications. Consequently, the estimate of the accuracy of prediction would be too reliant on the particular value selected. To ensure that this estimate was less dependent upon the particular value that was randomly selected to be used as the imputed value, 1000 repetitions were run for each variable to obtain an average number of correct re-classifications for each of the eight predictors in both the discriminant analysis and the logistic regression analysis.

Results

Using the linear discriminant function with nine predictor variables, 78 of the 99 individuals in the sample were correctly classified into the two groups: likely to experience a subsequent morbid event and unlikely to experience such an event. With the logistic regression analysis, 80 of the 99 individuals were correctly classified.

Overall, the results obtained by using the imputation techniques and comparing the re-classifications, as determined by the discriminant function and the logistic regression equation, with the actual group membership was encouraging. In general some, but not much, accuracy is lost when an original data value is replaced by an imputed value. The re-classification results are presented in Table 2.

For the discriminant analysis, the variable syncope was least affected by imputation, with 77, 78, and 76.7 individuals being correctly re-classified using mean, regression, and hot-deck techniques, respectively. (Recall, that the number of correct re-classifications using the hot-deck technique are averages of 1000 replications.) Mean substitution appeared to do slightly better than the other two techniques on most variables, particularly ventricular arrhythmia and heart sound S3. Overall, the average number of correctly re-classified individuals, averaging over all eight variables, was very similar for the three imputation techniques with mean substitution having an average number of correct re-classifications of 74.4, only slightly better than hot-deck imputation (73.1) and the regression method (72.0).

For the logistic regression analysis, the variable syncope was again the least affected by the imputation with numbers of correctly re-classified individuals of 78, 77, and 77.7, and systolic blood pressure, which correctly classified 77, 77, and 75.4 individuals also relatively unaffected by imputation. Overall, for the logistic regression analysis, mean substitution was again fairly consistent from variable to variable, although the hot-deck technique, with an average number of correct re-classifications of 73.9, was slightly better than mean substitution (73.5), and regression imputation (72.3).

Table 2. Numbers Of Correct Re-Classifications For Each Predictor Variable And Each Imputation Technique For Discriminant Analysis and Logistic Regression (n=99)

	Discriminant Analysis			Logistic Regression		
	Imputation Technique			Imputation Technique		
Variable Imputed	Mean	Regression	Hot-Deck*	Mean	Regression	Hot-Deck*
Age	74	76	72.4	75	77	72.4
Weight	74	72	72.4	73	73	71.4
Systolic Blood Pressure	74	74	73.8	77	77	75.4
White Blood Cell Count	71	71	70.2	72	70	70.1
Ventricular Arrhythmia	74	65	69.9	68	63	70.6
Syncope	77	78	76.7	78	77	77.7
Heart Sound S3	78	67	74.1	73	69	76.6
Aspirin	73	73	75.5	72	72	76.8
Mean of all predictors	74.4	72.0	73.1	73.5	72.3	73.9

* Values in this column represent the average number of correctly re-classified individuals for 1000 repetitions of a hot-deck imputation.

Discussion

The first phase of this research produced a linear discriminant function and a logistic regression model for classifying individuals as either likely or unlikely to have a subsequent morbid cardiac event after having first experienced an initial such event. Using a set of nine predictor variables, the discriminant function correctly classified 78 of the 99 individuals in the sample, while the logistic regression model classified 80 of the 99 individuals correctly. These numbers are not as high as we would have liked. However, given the relatively small sample size and the large number of variables that needed to be reduced to a manageable size, these results were the best that could be achieved. Given the results of previous research that identified potential predictors and the large number and variety of variables identified in that literature, it should not be surprising, perhaps, that any particular group of variables does not perform exceptionally well in predicting the outcome of interest.

Using either the discriminant function or the logistic regression model described above, the three imputation techniques, mean substitution, regression imputation, and hot-deck imputation, were compared to determine the extent to which replacing original data values with imputed values affected the number of correctly classified individuals. Overall, using an imputation technique to replace missing values in this application appeared to produce results which are comparable to those obtained using the actual data. Mean substitution was comparable to the hot-deck technique in the logistic regression analysis and slightly better than the other two techniques in the discriminant analysis. This result was somewhat surprising because of the general lack of trust that researchers appear to have in mean substitution for imputation. It was also somewhat satisfying, since mean substitution is a relatively easy technique to use and does not require sophisticated calculations, thus increasing the probability that it might actually be used in practice.

Perhaps it should not have been that surprising as well, since over 25 years ago, Chan & Dunn (1972) identified mean substitution as a preferred technique for imputation with discriminant analysis. One of the main criticisms of mean substitution is the fact that its use underestimates the variability in the variable being imputed. Regression imputation, on the other hand, does not have this same limitation, but these results indicate that regression imputation did not perform as well as either of the other two techniques in either the discriminant analysis or the logistic regression analysis, although the differences were not large. It was also a little surprising to observe that the regression technique did not perform better than the other two techniques, since this method is generally considered to be somewhat better in the sense that it incorporates other information about the individual in calculating the imputed value. This discrepancy might be explained by the fact that in classification, we are less concerned with predicting a specific value for an individual than we are with predicting that individual's group membership. Within each group are a variety of individuals who may possess a wide range of actual values on the criterion variable, which is much different from attempting to predict a specific outcome value (as is the case in multiple regression). Overall, it would appear that either mean substitution or hot-deck imputation would perform credibly in this application. Because mean substitution is easier to use than the hot-deck procedure, it would appear to be the better choice for practice.

There are, of course, limitations to this research. Our sample was relatively small for the number of predictors used. Larger samples with different predictors would likely produce at least a somewhat different discriminant function and/or logistic regression model. With different models, and different data, it is very likely that the number of correctly re-classified individuals would vary somewhat. With the relatively small differences among the imputation techniques, and between the two classification procedures, even slight differences in the re-classification results could lead to different conclusions than these. Furthermore, not all, nor necessarily even the best, imputation procedures were examined in this research. Choosing different techniques to investigate may also lead to different conclusions. Finally, it is uncertain how much our results are a function of the particular context, i.e., morbid cardiac events, within which we conducted this research, and how much would generally apply to other research scenarios. A different situation in which the predictor variables are very highly related to the outcomes of interest, resulting in extremely high numbers of correctly classified individuals in the original data, may be affected differently by imputation than was the case here.

At any rate, these results seem to indicate that using imputed values to replace missing values in classification models which have been previously derived from complete-case data can be a useful technique for making predictions we would otherwise be unable to make without re-calculating the models by leaving out the variables on which no data is available or having a series of models for use, each with a different combination of observed variables used as predictors. The ability to make such classifications with comparable accuracy, using a simple imputation technique such as mean substitution, would appear to be quite useful. By replacing missing values with the mean, thus being able to classify individuals who were previously unclassifiable using the same model, and to do so with a level of accuracy that is comparable to what would have been obtained had the values not been missing is a valuable tool. Furthermore the results were comparable regardless of whether the predictor variable being imputed was numerical or categorical.

References

Afifi, A. A. & Elasahoff, R. M. (1966). Missing observations in multivariate statistics I: Review of the literature, Journal of the American Statistical Association, 61, 595-604.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, Journal of the Royal Statistical Society, B22, 302-306.

Campbell, R. W. F., Murray, A., & Julian, D. G. (1981). Ventricular arrhythmias in the first 12 hours of acute myocardial infarction: Natural history study, British Heart Journal, 46, 351-357.

Chan, L. S. & Dunn, O. J. (1972). The treatment of missing values in discriminant analysis I: The sampling experiment, Journal of the American Statistical Association, 67, 473-477.

De Martini, M., Valentini, R., Cesana, B., Massari, F. M., Lettino, M., Pupilella, T., Ambrosini, F., Eriano, G., La Marchesina, U., & Lotto, A. (1990). Early and late prognosis in acute myocardial infarction: A retrospective study in patients admitted to the coronary care unit in the past 10 years, Italian Journal of Cardiology, 20, 215-226.

Gheorghiade, M., Anderson, J., Rosman, D. G., Lakier, J., Velardo, B., Goldberg, D., Friedman, A., Schultz, L., Tilley, B. & Goldstein, S. (1987). Risk identification at the time of admission to coronary care unit in patients with suspected myocardial infarction, American Heart Journal, 116, 1212-1217.

Haitovsky, Y. (1968). Missing data in regression analysis, Journal of the Royal Statistical Society, B30, 67-81.

Hartley, H. O. & Hocking, R. R. (1971). The analysis of incomplete data, Biometrics, 27, 783-808.

Kim, J. O. & Curry, J. (1977). The treatment of missing data in multivariate analysis, Sociological Methods and Research, 6, 215-241.

Kromrey, J. D. & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments, Educational and Psychological Measurement, 54, 573-593.

Little, R. J. A. & Rubin, D. B. (1987). Statistical analysis with missing data, New York, John Wiley and Sons.

Lown, B., Fakhro, A. M., Hood, W. B., & Thorn, G. W. (1967). The coronary care unit: New perspectives and directions, Journal of the American Medical Association, 199, 188-198.

Raymond, M. R. & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research, Educational and Psychological Measurement, 47, 13-26.

Rindskopf, D. (1992). A general approach to categorical data analysis with missing data, using generalized linear models with composite links, Psychometrika, 57, 29-42.

Roth, P. L. & Switzer III, F. S. (1995). A monte carlo analysis of missing data techniques in a HRM setting, Journal of Management, 21, 1003-1023.

Rubin, D. B. (1976). Inference and missing data, Biometrika, 63, 581-592.

Rubin, D. B. (1991). EM and beyond, Psychometrika, 56, 241-254.

Van Buuren, S. & van Ruckevorsel, J. L. A. (1992). Imputation of missing categorical data by maximizing internal consistency, Psychometrika, 57, 567-580.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028392

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Imputing Missing Values: the Effect on the Accuracy of Classification	
Author(s): Daniel J. Mundfrom, Alan J. Whitcomb	
Corporate Source: presentation at 1998 AERA Annual Meeting	Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☑	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Daniel J M	Printed Name/Position/Title: Daniel J. Mundfrom / Assoc. Professor	
Organization/Address: University of Northern Colorado Greeley, CO 80639	Telephone: 970-351-1669	FAX: 970-351-1622
	E-Mail Address: djmundf@unco.edu	Date: 4-13-98

(over)